A Sampling Theory for the Human Visual Sense

John Merchant
Honeywell Radiation Center, Boston, Massachusetts

16657 over

The subject of this paper is the nature of the sampling operation performed by the human visual sense, restricted to black and white, non-stereoscopic, photopic vision.

The hypothesis is presented that the human visual sense samples the "power" spectrum (in terms of spatial frequencies) of the input image, just as the aural sense samples the power spectrum of the input sound. The justification for this hypothesis

NOTE: This work was performed in part at Allied Research Associates, Concord, Mass, under NASA Contracts NASr-16 and NASw-535, and in part at the Honeywell Radiation Center.

is the fact that the sensitivity of the retina (except at the fovea) to form, or pattern, in the input image is very much poorer than is suggested by the corresponding upper cutoff spatial frequency of the retina. This property is characteristic of spectrum sensitive devices.

A physical model retina is described that could perform the hypothesized spectral sampling operation.

Author

INTRODUCTION

An image sensor may be defined as a device that can supply output information about an input image. Television and photographic cameras are examples of physical image sensors. The human visual sense is an example of a physiological image sensor.

The output signals of an image sensor can be utilized to infer certain properties of the input image. However, the sensor does not reveal everything about the input image, but merely performs a finite sampling operation on that image. This limitation is obvious in the case of a physical image sensor, such as a TV camera, and the nature of the sampling operation is also quite clear. The TV camera makes independent measurements of the average brightness of the input image over a finite number of resolution elements. It is less obvious that the human visual sense also performs a finite sampling operation, because of the tendency to identify the very strong subjective sensation of vision with objective reality. The nature of the sampling operation is also much more complex than in any physical image sensor.

THE GENERAL IMAGE SAMPLING OPERATION

The general image sampling operation can be expressed in terms of a function, b, defining the input image, and a sampling operator,

R, defining the sampling characteristics of the sensor. The input image function will be of two spatial dimensions and time, i.e., b (x, y; t), where b is the brightness at time, t, of that point in the input scene having angular coordinates (x, y) relative to axes fixed in the scene. The sampling operator will be a function of two spatial dimensions, i.e., R (ξ, η), where (ξ, η) are angular coordinates relative to axes fixed in the sensitive screen of the sensor. The result of the operation of R on b will be M, i.e.,

$$M(\xi, \eta; t) = R(\xi, \eta)\, b(x, y; t)$$

where M is the quantity actually measured at the point (ξ, η) of the sensitive screen at time, t.

In the equation above, (x, y) are the coordinates of the point of the input scene that is imaged onto the point (ξ, η) of the sensitive screen (Figure 1). The relationship between these coordinates is determined by the direction of pointing of the sensor relative to the input scene. This may be specified as the coordinates (X, Y) of that point in the input scene that is imaged onto the origin point of the sensitive screen. Then, approximately,

$$x = \xi + X$$

$$y = \eta + Y$$

The equation defining R may now be expressed in terms of (ξ, η), i.e.,

$$M(\xi, \eta; t) = R(\xi, \eta)\, b(X + \xi, Y + \eta; t)$$

In all practical cases the sampling operation may be regarded as being performed at a discrete set of points (ξ_i, η_j) $i = 1, 2, 3 \ldots\ldots$ $j = 1, 2, 3 \ldots\ldots$ rather than over a continuous range of values of (ξ, η). Similarly the temporal sampling operation may be considered as being performed at discrete instants of time rather than continuously. Various physical limitations will restrict the spatial and temporal bandwidths (in relation to the spectrum of the associated noise processes) so that the sampling theorem may be applied to show that the measurements actually made are equivalent to a discrete sampling. The total sampling performed by the sensor screen of the sensor will be a set of measurements, I, given by

$$I = \{M(\xi_i, \eta_j; t_k)\}$$

$$= \{R(\xi_i\, \eta_j)\, b(X_k + \xi, Y_k + \eta; t_k)\}$$

$$i = 1, 2, 3 \ldots\ldots \quad j = 1, 2, 3 \ldots\ldots \quad k = 1, 2, 3 \ldots\ldots$$

A general specification of the sampling operation performed by the sensor will also include a measurement of the direction of

pointing (X_k, Y_k) of the sensor during each 'frame-time', t_k.

In this paper, primary consideration will be given to the case in which the input image is time invariant (i. e., static). Then the result of the general sampling operation is the set

$$I = \{M(\xi_i, \eta_j)\}$$

$$= \{R(\xi_i, \eta_j)\, b(X + \xi, Y + \eta)\}$$

together with a measurement of (X, Y), the direction of pointing of the sensor.

PHYSICAL IMAGE SENSORS

In physical image sensors (i. e., cameras) the operator, $R(\xi_i, \eta_j)$ is simply an averaging operator. The set of samples derived by the sensor can be represented as

$$I = \left\{ \int\int_{\rho(\xi_i \eta_j)} b(X + \xi, Y + \eta)\, d\xi\, d\eta \right\}$$

$$i = 1, 2, 3 \ldots\ldots \quad j = 1, 2, 3 \ldots\ldots$$

where the integral is evaluated over the region $\rho(\xi_i, \eta_j)$, corresponding to the area of the i, j^{th} resolution element of the sensor.

(The spatial averaging will generally not be quite as simple as indicated here but the differences involved are not significant in the present context.)

In photography this set of samples (I) is very large. Each member of the set corresponds to an element of grain in the developed film. There is no measurement of the direction of pointing of the camera (X, Y).

In television, the instantaneous measurement is of the average brightness of the input image at one point only -- the point on the photocathode covered, at any instant, by the scanning electron beam. That is, the set I contains only one sample, i.e.,

$$I = \int_{\rho} \int_{(0,0)} b\,(X + \xi,\ Y + \eta)\ d\,\xi\ d\,\eta$$

The sampling operator R (ξ, η) exists at one central point only ($\xi = 0$, $\eta = 0$). There is a measurement of the instantaneous position (X, Y) of the scanning electron beam. This information is contained in the sync pulses. The sampling operation performed by a TV camera over one frame time is essentially the same as in the case of a photographic camera. There is no measurement of the direction of pointing (X, Y) of the television camera.

In physical image sensors the samples taken are of the most elementary nature and thus the entire sampling operation can be

specified simply in terms of the sampling density, i.e., the number of independent samples obtained per unit angular area of the sensitive screen. In physical image sensors, the sampling density is usually uniform over the field of view of the sensor.

The sampling density determines the resolution (or acuity) of the sensor. Resolution may be measured by determining the smallest angular size of a certain test object, or pattern, that can just be resolved or identified by the sensor. Two test objects that can be used are a grating of equal dark and light bars and an alphabetic character. The corresponding acuities may be termed the grating and letter acuities. It is obvious that for a physical image sensor the grating acuity and the letter acuity are not independent, but are both measures of the same thing -- the sampling density. The numerical difference between the grating and letter acuities (approximately a factor of 3) is simply a function of the relative complexity of the two patterns.

THE HUMAN VISUAL SENSE

The sampling operation performed by the human visual sense can be exactly defined and measured only in terms of psychophysical experiments. Nevertheless, the characteristics of the various component parts of the system, e.g., the refractive surfaces, the retina, the optic nerve -- will determine, to a considerable extent,

the characteristics of the visual sense. Much can be inferred, therefore, about the visual sense from these structural characteristics.

The primary effect of the refractive surfaces is to limit the spatial frequency content of the retinal image. Distortion of the retinal image is not, in itself, a limitation since it can be, and presumably is, corrected for when the retinal signals are interpreted by the brain.

The initial sampling of the retinal image is performed by the rods and cones. These cells absorb light energy and generate neural signals. However, the set of signals produced by the photodetectors is not the set actually transmitted to the brain over the optic nerve. The number of individual fibers in the optic nerve is only about 1% of the total number of photoreceptors, so that the initial set of neural signals must be compressed to a much smaller set prior to transmission. This is accomplished by neural networks within the retina. The nature of the individual samples taken by the human visual sense depends, to a large extent, on the way in which the initial set of signals is compressed into the set actually transmitted to the brain. The nature of the sampling operation performed by the retina is also related to the density of samples taken by the photoreceptors (i.e., the number of samples per unit angular area) and on the

density of samples transmitted from the retina to the brain. The former density may be associated with the photoreceptor density, the latter with the density of the terminal points, in the retina, of the optic nerve fibers.

The distribution density function (i.e., the number per unit area of the retina as a function of angular position on the retina) of the rods and cones is shown in Figure 2[1]. The combined density of rods and cones has a broad maximum around the central foveal axis of the retina.

The distribution density function of the terminal points of the optic nerve fibers can be estimated. There is an approximately one-one relationship at the fovea between cones and optic nerve fibers. Thus the fiber density function must have approximately the same value at the fovea as the cone density function. Over the whole retina however, there are about seven times as many cones as optic nerve fibers so that the fiber density function must be considerably less than the cone density at all other parts of the retina. The fiber density function estimated in this way is shown in Figure 3.

The estimated fiber density function indicates that, in marked contrast to physical image sensors, the sampling density function in human vision is extremely non-uniform over the field of view. However, the immediate subjective sensation of vision and the

observed general visual capability of the human subject are suggestive of a fairly uniform, wide-angle, high resolution sensing of the environment by the human visual sense. An explanation of this paradox is suggested by analogy with those physical image sensors that also generate detailed images with a very narrow sampling beam -- for example, PPI radar and television. In television, the primary sampling operation is performed at one point only, where the electron beam hits the photocathode. Nevertheless a wide-angle, detailed image is built up from the information supplied by this narrow sampling beam. In general, the conditions under which a detailed image can be built up from the results of a narrow beam sampling operation are:

(a) The narrow sampling beam must be made to perform a precisely controlled scanning action over the input image

(b) The instantaneous direction (X, Y) of the sampling operator must be measured and this information used in the generation of the output image (cf the sync information in TV).

The sampling density function in human vision is intermediate, in general form, between the instantaneous sampling density

function in TV, where the conditions (a) and (b) hold, and the uniform sampling density function of a movie camera, for which the conditions do not apply. It is possible therefore that, to some extent at least, the human visual sense also depends on a controlled scanning action to generate a wide-angle detailed image. Studies of the oculomotor system of the eye have shown that the pointing of the eyeball is very rapid and precise, in contrast to the relatively casual operation of pointing a camera. This suggests that condition (a) above is satisfied. There is evidence to show that condition (b) is also satisfied. Studies of the oculomotor system of persons with defective spatial perception have lead to the conclusion that the afferent link in the oculomotor system is essential for proper spatial perception[2].

Psychophysical measurements of visual acuity, as a function of angular position on the retina are qualitatively consistent with the form of the sampling density function inferred from consideration of the retinal structure. Acuity is highest at the fovea and falls off rapidly away from the fovea. A measurable drop of acuity has been observed only 3.5 minutes of arc from the center of the fovea.[3]

Visual acuity may be measured in various ways depending on the test object, or pattern, used. In the case of physical image sensors these various measures of acuity -- e.g., grating

acuity and letter acuity -- are not independent but are in a fixed ratio to each other which is independent of the sensor or of the resolution. However, this is not the case with human vision. The ratio of the grating acuity to the letter acuity is not constant over the retina. At the fovea the ratio has about the same value as in physical image sensors but in the peripheral retina, letter acuity is relatively much poorer than grating acuity. The variation in the ratio of the two acuities is very considerable and can be demonstrated qualitatively as shown in Figure 4.

It is impossible to account for the variation over the retina of the ratio of the acuities with a sampling model in which -- as in all physical image sensors -- each sample is a simple spatial average of the brightness function taken over the area of a resolution element. Thus it must be concluded that the human visual sense differs from physical image sensors, not only by virtue of the very non-uniform sampling density function of the retina, but also because most of the retinal samples are different, in nature, to the samples taken by physical image sensors.

THE RETINAL SAMPLING OPERATOR

The variation over the retina of the ratio of the grating and letter acuities has been cited as an anomalous characteristic

of the human visual sense. Another anomaly is illustrated in Figure 5. This shows that a single, sharp, black line on a plain white background is clearly identifiable as such using the peripheral retina. However when additional lines are added, the individual identity of the original line is lost. The sensation is then not of the individual lines but of the network. This effect is not observed in foveal vision, or in any physical image sensor. For example, the visibility of the line shown in Figure 5a would not be affected, in a photograph, by the presence of the adjacent lines shown in Figure 5b.

Analogs of these anomalous characteristics of vision can be found in the human aural sense. Normally, the ear can detect a 10 kc audio sine wave -- indicating a temporal resolution capability of 100 μs. It is not possible, however, to resolve Morse code characters unless they extend over periods of the order of 100 ms -- a thousand times greater than the sinusoidal temporal resolution of the ear. However, a microphone and oscilloscope display system that had sufficient 'resolution' (i.e., bandwidth) to detect a 10 kc sine wave would be able to resolve Morse code characters having a temporal duration of, say, 500 μs -- i.e., only five times the sinusoidal temporal resolution of the system. A single 100 μs sonic pulse may be detected by the ear as a sharp "click". However, if other similar temporarily adjacent pulses are mixed at random with it, the

ear is no longer able to detect the original pulse as a unique entity. The sensation becomes that of a composite hiss. Again, however, the microphone and oscilloscope system would be quite unaffected in its ability to resolve the original pulse by the presence of other temporally adjacent pulses.

These characteristics of the aural sense are anomalous in that they are inconsistent with a simple temporal-averaging sampling model. That is, the type of sampling performed by a microphone/oscilloscope system at intervals of one-half of the reciprocal of the cutoff frequency of the system. The characteristics of the aural sense are much closer to those of a physical spectrum analyser. A spectrum analyser is a device which effectively samples the modulus of the Fourier Transform of the input signal, rather than the input signal itself. A spectrum analyser could be designed so that, like the ear, it might have a resolution of 100 μs (that is, be able to detect a 10 kc note) but be unable to resolve Morse code characters until they extended over a period of 100 ms.

The anomalous characteristics of the aural senses that have been cited can be explained, qualitatively, by a temporal spectral theory of aural sensing. This asserts that aural sense samples the modulus $[P(\omega, t)]$ of the Fourier Transform of the input stimulus function p (t) evaluated over a period of time of the order of 100 ms:

$$P(\omega, t) = \left| \int_{t-0.05}^{t+0.05} p(\tau)\, e^{-i\omega\tau}\, d\tau \right|^2$$

The hypothesis is now presented that analogously, a spatial spectral theory of visual sensing may be able to account for the anomalous characteristics of the visual sense. According to this theory the sampling operation of the human visual sense would be performed, not on the input image b (x, y), but on the Wiener Spectrum[4] ("power" spectrum) of this function, i. e.,

$$B(u, v; x, y) = \left| \iint_{W(x,y)} b(l, m)\, e^{-i(ul + vm)}\, dl\, dm \right|^2$$

where the integral is evaluated over a region W. This sampling model provides a qualitative explanation for sensitivity to high spatial frequencies without a corresponding sensitivity to pattern.

SPATIAL SPECTRAL HYPOTHESIS

The sampling operation for any image sensor has been expressed in terms of a sampling operator R (ξ η) representing the actual operation performed on the input function [b (xy)] as a function of position (ξ, η) on the sensitive screen. Thus M (ξ_i, η_j) is the value of the sample taken at (ξ_i, η_j) where

$$M(\xi_i, \eta_j) = R(\xi_i, \eta_j)\, b(X + \xi, Y + \eta)$$

and where (X, Y) is the direction of pointing of the sensor.

The spatial spectral hypothesis may be formally expressed as follows:

$$R(\xi_i\, \eta_j)\, b(X + \xi, Y + \eta) = F(\xi_i\, \eta_j)\, B(u, v; \xi_i\, \eta_j)$$

where F is a sampling operator acting on the Wiener Spectrum of the input function b (x, y) i. e. ,

$$B(u, v; \xi_i\, \eta_j) = \left| \iint_{W(\xi_i\, \eta_j)} b(X + \xi, Y + \eta)\, e^{-i(u\xi + v\eta)}\, d\xi\, d\eta \right|^2$$

The function B (u, v; $\xi_i\, \eta_j$) is defined for spatial frequencies up to a cutoff spatial frequency $\Omega(\xi_i\, \eta_j)$.

The result of the total sampling operation is the set of measurements I;

$$I = \{F(\xi_i\, \eta_j)\, B(u, v; \xi_i\, \eta_j)\}$$

$$i = 1, 2, 3 \ldots\ldots \quad j = 1, 2, 3 \ldots\ldots$$

The basic parameters of this model are Ω, W, and F. Ω is the highest spatial frequency to which the retina can respond. It may be associated with the grating acuity and the photoreceptor density. W is the angular size of the region of the retina over which the Wiener Spectrum is computed. Within this area there is no sensitivity to pattern, only to the spatial frequency content of the input image. W may be associated with the letter acuity and the optic nerve fiber density. At the fovea, W would be approximately equal to the size of the elementary resolution elements, that is about one minute of arc ($W \sim \pi/\Omega$), and the spectral sampling operation would degenerate to a simple spatial averaging. In the peripheral retina, W would be considerably greater than the size of a resolution element ($W \gg \pi/\Omega$) and might typically be of the order of a few degrees. The result of the operation, F, is a set of weighted spectral averages, e.g.,

$$F(\xi_i \eta_j)\, B(u, v; \xi_i \eta_j) = \{ \int_0^\Omega \int_0^\Omega B(u, v; \xi_i \eta_j)\, K_p(u, v)\, dudv \}$$

where K_p, $p = 1, 2, 3 \ldots\ldots$, is a set of spectral weighting functions which specify the operator F.

This specification of F will define what properties of the Wiener Spectrum of the input image are sensed. A similar situation exists in color vision. The tristimulus theory of color sensing asserts that the photometric measurements made by the retina

are of three weighted averages (M_i, i = 1, 2, 3) of the input spectrum of light energy H (λ), i.e.,

$$M_i = \int_{\lambda_1}^{\lambda_2} H(\lambda)\, J_i(\lambda)\, d\lambda$$

$$i = 1, 2, 3.$$

where $J_1(\lambda)$, $J_2(\lambda)$, $J_3(\lambda)$ are three spectral weighting functions which define the tristimulus sampling model.

The spatial spectral theory of visual sensing has been presented in terms of parameter functions Ω, W, and K_p, p = 1, 2, 3 ..... This is as far as it is possible to go in a theoretical formulation of this sampling model. The primary justification for the theory, as it stands, is that it can account, in qualitative terms, for the two anomalous characteristics of peripheral vision considered earlier.

The spectral theory may be tested by determining the extent to which it is possible to fit the psychophysical characteristics of vision within the theoretical framework that has been presented. This will involve appropriate selection of the parameter functions;

(1) $\Omega(\xi, \eta)$ - the upper spatial cutoff frequency, as a function of retinal position

(2) $W(\xi, \eta)$ - the area over which the Wiener Spectrum is computed, as a function of retinal position

(3) $K_p(u, v)$ - the spectral weighting functions.

An exact model of the human visual sense must also include a specification of the associated noise processes. This aspect of the problem has not been considered here. However, an approximate description of a sensor may be given, in many practical cases, without reference to noise. For example, when the rate of cutoff beyond the nominal (3 db) cutoff frequency of a sensor is very steep, the effective information bandwidth may be almost independent of the noise level over the practical operating range of the sensor.

Ideally, a sampling and noise model for the human visual sense should be able to account for all the observed characteristics of vision and also all the sensing characteristics implied by the model should be observed in the human visual sense.

PHYSICAL MODEL

In order to show that the proposed spatial spectral sampling operation is at least physically realisable, a physical model retina will be described which would perform the same sampling operation as that proposed as a model for the human visual sense.

The model retina consists of a set of receptor units. A typical receptor unit, having a field of view, $W(\xi_i\ \eta_j)$, is shown in Figure 6, located at the point $(\xi_i\ \eta_j)$ of the retina. Similar units are located at other points of the retina so that the whole field of view is covered. Each receptor unit consists of an objective lens, a special reticle system, and a single photodetector with its associated electronics. The input image to a receptor unit is focussed onto the reticle plane. The quality of the image is designed to be such that the upper cutoff spatial frequency of the image is $\Omega(\xi_i,\ \eta_j)$.

The reticle consists of a transparency, the transmission factor of which is a random function of position, band limited to a cutoff spatial frequency $\Omega(\xi_i\ \eta_j)$ and having a specific Wiener Spectrum $K_1(u,\ v)$. (The reticle is a two dimensional analog of a noise function having a specific spectral distribution -- i.e., filtered white noise.) In the operation of the sensor, the reticle is abruptly changed for another reticle at a rate of f_r times per second. Each reticle is different, but all have the properties listed above. The mean square value of the fluctuating component of the photocell output is measured while the reticles are being changed in this fashion. After this has been done, the whole process is repeated for different reticle functions $K_2(u,\ v)$, $K_3(u,\ v)$ ..... etc. Appendix B shows that the mean square value of the fluctuating component of the photodetector output, in each

case, is proportional to a weighted average of the Wiener Spectrum of the input image where the weighting function is the corresponding function K_1 (u, v), K_2 (u, v), K_3 (u, v) ..... etc. That is, the detector circuit shown in Figure 6 measures

$$\int_0^{\Omega}\int_0^{\Omega} B\,(u,\ v;\ \xi_i\ \eta_j)\ K_p\,(u,\ v)\ d\,u\,d\,v$$

$$p = 1,\ 2,\ 3\ .....\ \text{etc.}$$

$$\Omega = \Omega\ (\xi_i,\ \eta_j)$$

where

$$B\,(u,\ v;\ \xi_i\ \eta_j) = \left| \int\int_{W\,(\xi_i\ \eta_j)} b\,(X + \xi,\ Y + \eta)\ e^{-i(u\xi\ +\ v\eta)}\ d\,\xi\ d\,\eta \right|^2$$

These measurements are exactly those proposed in the spatial spectral model of the human visual sense.

It is very probable that the human retina samples the average value of the brightness function over the area W as well as sampling certain weighted averages of the Wiener Spectrum within this area. That is, a measurement of the quantity

$$B(0, 0; \xi_i, \eta_j) = \left| \int\int_{W(\xi_i, \eta_j)} b(X + \xi, Y + \eta)\, d\xi\, d\eta \right|^2$$

probably forms part of the retinal sampling operation. This particular sample can be included within the framework of the proposed spatial spectral model by assigning one of the spectral weighting functions, $K_p(u, v)$, as the delta function. In the physical model retina this spatial average, $B(0, 0; \xi_i \eta_j)$, is measured as the square of the average (DC) value of the detector output as the reticles are charged.

The image information supplied by the model retina consists of a low pass filtered version of the input image (in the form of a set of spatial averages over the resolution areas $W(\xi_i \eta_j)$) together with certain measurements of the average Wiener spectrum of the rejected high pass portion of the input image (Figure 7). Since the Wiener averages are taken over the resolution areas W, the high pass Wiener signal requires a bandwidth only of the same order as the low pass channel. This method of transmission makes it possible, therefore, to transmit high (spatial) frequency image detail using a smaller total channel capacity (bandwidth) than would otherwise be necessary. The penalty paid for this gain is a degree of uncertainty, or ambiguity, in the received image.

When the input image is very simple there is little ambiguity. For example, consider the image shown in Figure 5a which is of a single thin, sharp line on a plain background. The direct low pass signal will give the general form of the image as a thick blurred line, and the high pass Wiener signal will clearly indicate that the line is actually thin and sharp. When additional lines are added to the single line to make a more complex input image, as in Figure 5b, the spectral transmission method becomes ambiguous. The direct low pass signal will be a much poorer reproduction of the form of the input image because the blurred image of one line will overlap onto that of an adjacent line. The high pass Wiener signal cannot resolve ambiguities of form within the low pass resolution elements. All that can be determined about the input image, therefore, is that it is a network of thin sharp lines. The actual details of structure cannot be resolved.

According to the spatial spectral theory, the physical model retina, with appropriate functions for the parameters Ω, W, and K_p, p = 1, 2, 3 ....., makes exactly the same measurements on the input image as does the human retina and this is all that is intended for the model retina. It is not meant to reflect the actual construction, or method of operation, of the eye. However, there are certain other, unintentional, similarities between the physical model retina and the human retina which are worthy of passing note.

The physical model retina, like the human retina, has a sensitivity to rate of change between frames. That is, the sampling operator R has a capability for temporal sensing of the second type (see Appendix A). With a static image, the photocell output in the period of $1/f_r$ seconds that any one reticle is in position, would be a constant (Figure 8). However, any movement in the image during this period gives rise to fluctuations in the detector output as the moving element in the image moves over the variable transmission reticle. Thus, fluctuations in the photocell output during the $1/f_r$ period are a certain indication of image motion. Obviously, this motion sensing mechanism could not operate during the short time that a reticle was being changed or while the whole retina was in motion. The photocell output will fluctuate under these conditions whether or not there is any image motion, so that detector fluctuations cannot then serve as a certain indication of image motion. There is evidence to suggest that in the human retina, the motion sensing mechanism is also "turned off" during saccades[5] or when the gaze is moved from one point to another. For example, it is not possible to detect any motion in one's own eyes when looking at them in a mirror.

There are other features of the human retinal system that are somewhat similar to those of the physical model retina and these might provide the basis for a spectral sensing mechanism. The random distribution of photoceptors might correspond to the random noise pattern of the changing reticles and the nystagmus

to the changing of the reticles. If this were so, it would follow that the artificial stabilization of an image on the human retina would immediately result in the loss, by the brain, of the postulated spatial spectral information.

IMPLICATIONS OF A SAMPLING MODEL OF THE HUMAN VISUAL SENSE

Irrespective of the validity, or otherwise, of the particular sampling model that has been proposed here (the spatial spectral hypothesis), the implications and applications of the establishment of an accurate sampling model for the human visual sense may be considered.

In many respects, a microphone at least equals the performance of the human aural sense as a sound sensor. A suitable microphone, amplifier, recording device, and loudspeaker system can almost completely satisfy the aural sense. However, the human visual sense is vastly superior to its electronic analog. The visual sense supplies far more information about the input image than a 500 line television system which fails by a wide margin to satisfy the visual sense.

There are three possible explanations for this difference between the visual sense and television.

1) The amount of information transmitted per second from the eye to the brain might be much greater than the amount supplied per second by a television camera.

2) The special data processing capability of the brain might be responsible for the difference.

3) The type of information sensed by the eye might be more useful to the human subject than the type of information sensed by a television camera.

The first possibility can be rejected because the sample capacity, and the associated noise levels, are of the same order of magnitude in the two image sensors. There can be no more than 10^6 samples per "frame" in the visual sense (this is the number of optic nerve fibers), and there are 0.25×10^6 samples in a 500 line television picture.

The second possibility must be rejected because subsequent data processing, no matter how sophisticated, cannot create information that is absent in the sensor output signals.

Therefore, the superior performance of the visual sense is due, in large measure, to the type of information to which its limited channel (sample) capacity is devoted.

The type of information that is sensed is determined by the nature of the samples taken. Except for the small foveal area, the samples taken by the visual sense are of a different nature to those taken by physical image sensors. In spite of the heavy concentration of sensory capability at the fovea, foveal vision occupies only a small fraction of the total channel capacity of the sense because most of the optic nerve fibers serve the peripheral retina. It follows, therefore, that the nature of the peripheral sampling operation may be an important factor relating to the superior performance of the human visual sense.

The elucidation of the nature of the sampling operation performed by the visual sense will help to explain, in physical terms, many of the psychological and psychophysical properties of vision. It will also be of use in visibility analysis, for example, to predict search times, design for maximum or minimum visibility, etc.

There may also be applications of a visual sampling model in physical systems. For example, it might be possible to construct a television system that could perform the same sampling operation as the human visual system and that could display the resulting data in such a way that it could be properly absorbed by the human subject. A television system of this type -- which would necessarily be a one-viewer-per-camera system, would be able to provide a very high degree of remote visual capability

with a video bandwidth no greater than that of a conventional 500 line system. It would be a television analog of the vocoder system of sound transmission.

CONCLUSION

A sampling model for the human visual sense has been proposed in which each sample taken of the input image is a weighted average of the Wiener spectrum of that image, evaluated over an area W for spatial frequencies up to Ω, where W and Ω are both functions of the position on the retina.

At the fovea, the size of the area W is of the same order as the period $2\pi/\Omega$ of the upper cutoff spatial frequency, and the spectral sampling become indistinguishable from a direct sampling of the input image at the resolution intervals of π/Ω. At the other parts of the retina, the size of the area W is greater than the corresponding period, $2\pi/\Omega$, of the upper cutoff spatial frequency. The sampling model thereby allows for sensitivity in the peripheral retina to relatively high spatial frequencies within the area W, but with a sensitivity to pattern or form to a lower resolution as defined by the size of the area W.

A physical model retina has been described which would perform the same sampling operation as that proposed for the human retina.

The spatial spectral theory has been presented in terms of parameter functions Ω, W, and K_p, p = 1, 2, 3 ..... Its validity depends on the extent to which it may prove possible to select actual functions for these parameters to account for the psychophysical characteristics of vision.

The primary justification for the theory, as it stands, is that it can account, in qualitative terms, for two anomalous characteristics of peripheral vision that have been described. The theory uses concepts of Fourier analysis in an attempt to unify various characteristics of human vision, and may therefore be regarded as a simplifying generalization. In view of the very complex nature of the human visual sense, the theory is unlikely to be exactly true, but may, nevertheless, be useful as a simple approximation.

The establishment of an accurate sampling model for the visual sense will help to account for the superior capability of this sense. The model might find practical application in a television system designed to perform the same sampling operation as the visual sense while displaying the resulting data in a form that could be properly absorbed by a human subject. This television system will provide a very high degree of remote visual capability with the same video bandwidth required by conventional television.

APPENDIX A
SAMPLING IN THE TIME DOMAIN

The general sampling operation performed by an image sensor can be expressed symbolically as

$$\{M(\xi_i, \eta_j; t_k)\} = \{R(\xi_i, \eta_j)\, b(X_k + \xi, Y_k + \eta; t)\}$$

Temporal information may be derived, within this sample set, in two ways. In the first, each sample, taken at t_k, $k = 1, 2, 3 \ldots\ldots$ will be a temporal average of the input image over the sample interval, i.e.,

$$\{M(\xi_i, \eta_j, t_k)\} = \{R(\xi_i \eta_j) \int_{t_k}^{t_{k+1}} b(X_k + \xi, Y_k + \eta, t)\, d\, t\}$$

Temporal information is derived by comparing the sample sets at the various times t_k, t_{k+1}, t_{k+2}, etc. This is the sort of temporal sampling performed by a movie camera. In the second method of temporal sampling the operator R has a specific sensitivity to rate of change within each frame period. A movie camera does not perform temporal sampling of the second type. In human vision it would appear that there is temporal sampling of the first type, at a frame rate of about 5 cps, together with some temporal sampling of the second type that, presumably, prevents the occurrence of stroboscopic sensations.

APPENDIX B

MEASUREMENT OF THE WIENER SPECTRUM

The transmission factor of the reticle in the ij^{th} receptor unit of the physical model retina is a random function, $t(\xi, \eta)$, of position having a prescribed Wiener Spectrum, $K_p(u, v)$, band limited to an upper cut of spatial frequency $\Omega(\xi_i \eta_j)$, and with a field of view $W(\xi_i, \eta_j)$ centered at $(\xi_i \eta_j)$. Let the field of view W be square, $\lambda \times \lambda$, where

$$\frac{\Omega}{2\pi} = \frac{n}{\lambda}$$

Then the function t may be represented as a Fourier Series,

$$t(\xi, \eta) = a_{00} + \sum_{1}^{n} \sum_{1}^{n} a_{rs} \sin\left(\frac{2\pi}{\lambda}(r\xi + s\eta) + \alpha_{rs}\right)$$

The numbers a_{rs} are specified by the prescribed Wiener Spectrum $K_p(u, v)$ i.e.,

$$a_{rs}^2 = K_p\left(\frac{2\pi r}{\lambda}, \frac{2\pi s}{\lambda}\right)$$

The phase angles α_{rs} are random, and distributed uniformly in the range 0-2 π.

The image $b(X + \xi, Y + \eta)$ that is formed on the reticle may also be represented by a Fourier Series;

$$b(X+\xi,\ Y+\eta) = b_{00} + \sum_{1}^{n}\sum_{1}^{n} b_{rs} \sin\left(\frac{2\pi}{\lambda}(r\xi + s\eta) + \beta_{rs}\right)$$

The $2n^2$ numbers b_{rs} and β_{rs} define a general bandlimited function over the area W. The Wiener Spectrum of the image is given by

$$b_{rs}^2 = B\left(\frac{2\pi r}{\lambda},\ \frac{2\pi s}{\lambda};\ \xi_i,\ \eta_j\right)$$

The electrical output of the photodetector is proportional to the total luminous flux passed by the reticle, i.e., to

$$P = \int_0^{\lambda}\int_0^{\lambda} b(X+\xi,\ Y+\eta)\, t(\xi,\ \eta)\, d\xi\, d\eta$$

Using the orthogonal properties of the sine function;

$$P = \lambda^2 a_{00} b_{00} + \int_0^{\lambda}\int_0^{\lambda} \sum_{1}^{n}\sum_{1}^{n} a_{rs} b_{rs} \sin\left(\frac{2\pi}{\lambda}(r\xi + s\eta) + \alpha_{rs}\right)$$

$$\sin\left(\frac{2\pi}{\lambda}(r\xi + s\eta) + \beta_{rs}\right) d\xi\, d\eta$$

$$= \lambda^2 a_{00} b_{00} + \frac{1}{2}\int_0^{\lambda}\int_0^{\lambda} \sum_{1}^{n}\sum_{1}^{n} a_{rs} b_{rs} \left\{\cos(\alpha_{rs} - \beta_{rs})\right.$$

$$\left. -\cos\left(\frac{4\pi}{\lambda}(r\xi + s\eta) + \alpha_{rs} + \beta_{rs}\right)\right\} d\xi\, d\eta$$

$$= \lambda^2 a_{00} b_{00} + \frac{\lambda^2}{2} \sum_{1}^{n} \sum_{1}^{n} a_{rs} b_{rs} \cos (\alpha_{rs} - \beta_{rs})$$

Since the n^2 phase angles α_{rs} are random, with a uniform distribution in 0-2 π, the quantity $a_{rs} b_{rs} \cos (\alpha_{rs} - \beta_{rs})$ is randomly distributed about a zero mean with a standard deviation of 1/2 $(a_{rs} b_{rs})$. Therefore the mean square value of the fluctuating component of the photodetector output is proportional to

$$\text{Variance } (P - \lambda^2 a_{00} b_{00}) = \frac{\lambda^4}{8} \sum_{1}^{n} \sum_{1}^{n} a_{rs}^2 b_{rs}^2$$

$$= \frac{\lambda^4}{8} \sum_{1}^{n} \sum_{1}^{n} K_p \left(\frac{2 \pi r}{\lambda}, \frac{2 \pi s}{\lambda} \right) B \left(\frac{2 \pi r}{\lambda}, \frac{2 \pi s}{\lambda}; \xi_i, \eta_j \right)$$

$$\approx \frac{\lambda^6}{32 \pi^2} \int_0^{\Omega} \int_0^{\Omega} K_p (u, v) B (u, v; \xi_i \eta_j) \, d u \, d v$$

Thus the mean square value of the fluctuating component of the output of each receptor unit of the model retina is proportional to a weighted average of the Wiener Spectrum of the input image, where the weighting function is the Wiener Spectrum of the reticle function.

REFERENCES

1. M. H. Pirenne, Vision and the Eye (Chapman and Hall, London, 1948).

2. L. I. Leushina and Y. P. Kok, Conference on the Problems of Spatial Perception and Spatial Concepts, Leningrad, May 1959. NASA Technical Translation NASA TTF-164.

3. L. A. Jones and G. Higgins, J. Opt. Soc. Am. 37, 217 (1947).

4. R. Clark Jones, J. Opt. Soc. Am. 45, 799 (1955).

5. R. W. Ditchburn, Optica Acta 1, 171 (1955).

FIGURES

Figure 1 The point P in the scene is imaged at P' on the sensor screen. Relative to scene axes the angular coordinates of P are (x, y) and relative to screen axes the angular coordinates of P' are (ξ, η). The direction of pointing of the sensor is defined by the coordinates (X, Y).

Figure 2 Distribution of rods and cones in the human retina. (See footnote 1)

Figure 3 Estimated fiber density function, based on cone density equal to fiber density at fovea and total number of cones equal to seven times total number of fibers.

Figure 4 When the smaller pattern is viewed foveally so that the grating is just resolvable the letters SHON are identifiable. When the larger pattern is viewed peripherally so that the grating is just resolvable the letters are unidentifiable. The ratio of the grating separation to the letter size is the same in both patterns.

Figure 5 The single line in figure 5a is clearly identifiable with peripheral vision. However the identity of this line is lost when other lines are added to it (figure 5b).

Figure 6 Receptor unit of physical model retina.

Figure 7 Spatial spectral method of image transmission.

Figure 8 Receptor-unit protocell output as a function of time with and without image motion.

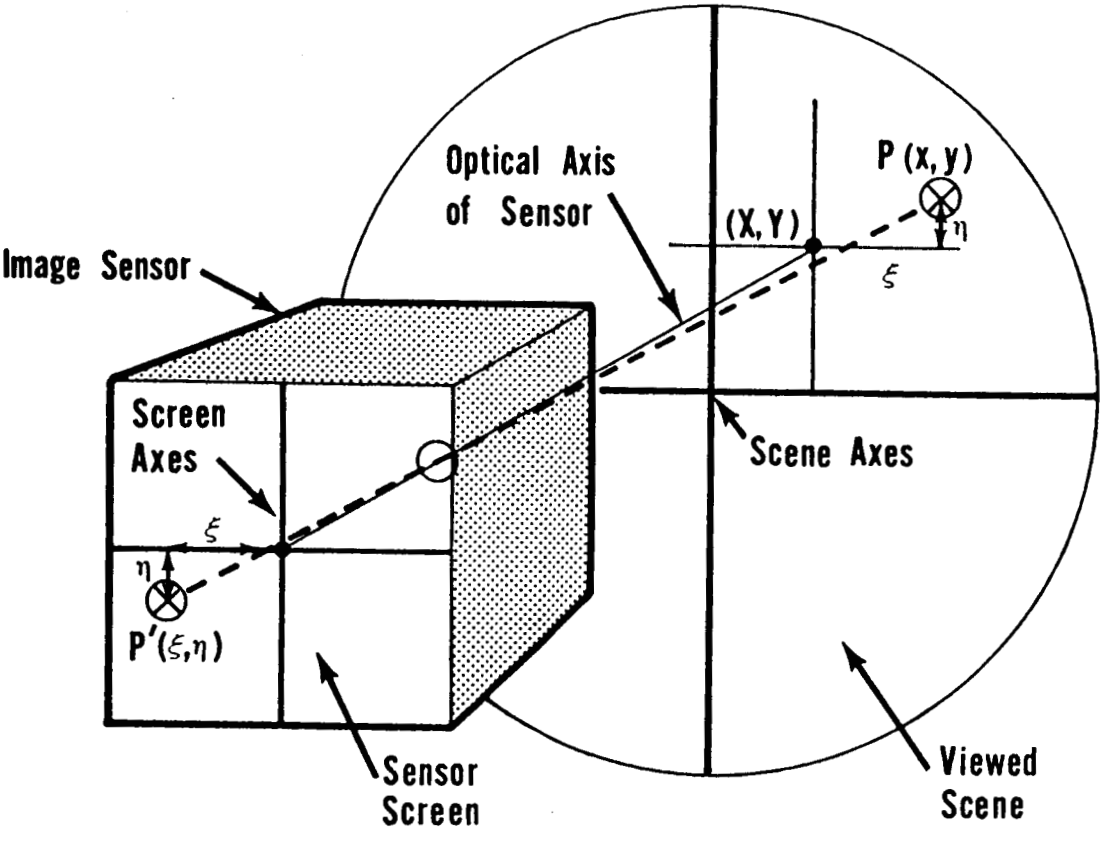

Figure 1 The point P in the scene is imaged at P' on the sensor screen. Relative to scene axes the angular coordinates of P are (x, y) and relative to screen axes the angular coordinates of P' are (ξ, η). The directing of pointing of the sensor is defined by the coordinates (X, Y).

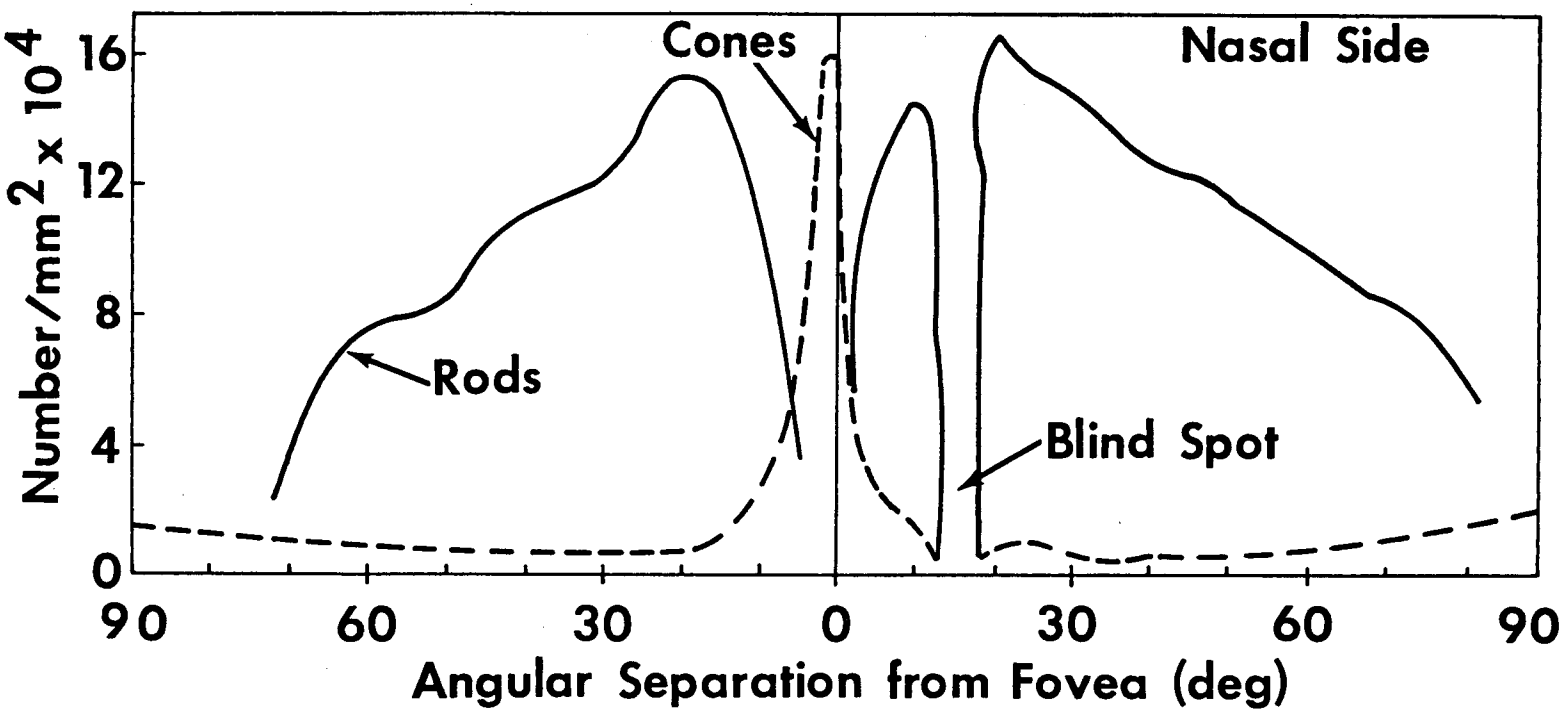

Figure 2

Distribution of rods and cones in the human retina (see footnote 1)

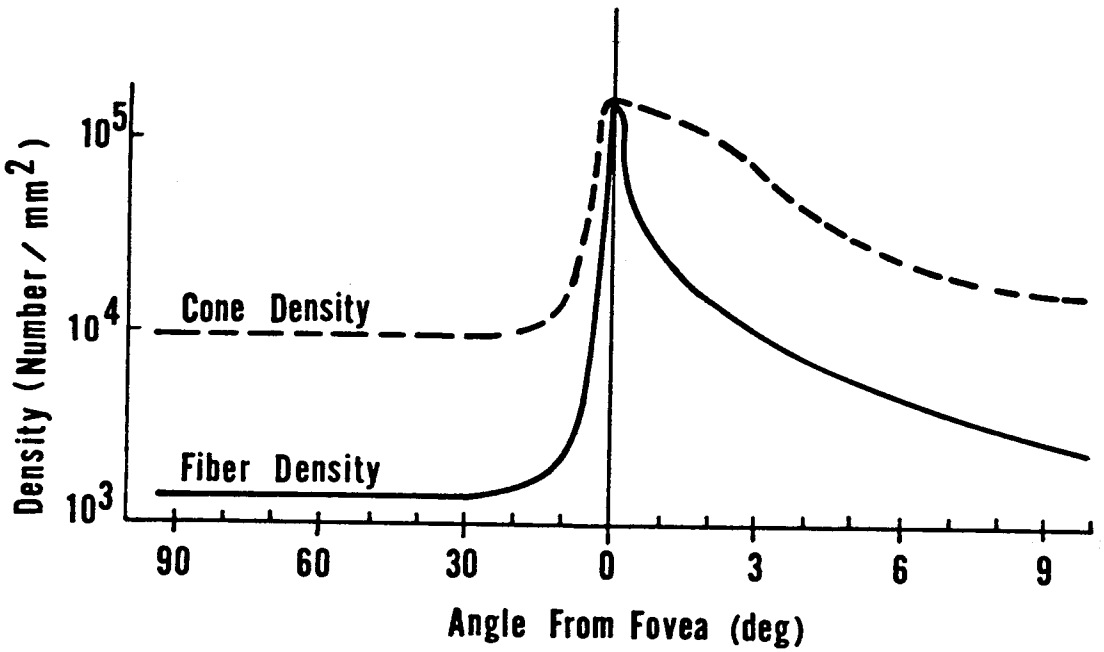

Figure 3

Estimated fiber density function, based on cone density equal to fiber density at fovea and total number of cones equal to seven times total number of fibers.

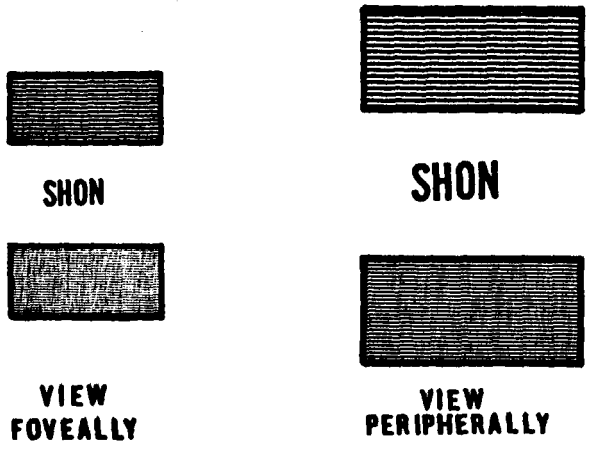

Figure 4 When the smaller pattern is viewed foveally so that the grating is just resolvable the letters SHON are identifiable. When the larger pattern is viewed peripherally so that the grating is just resolvable the letters are unidentifiable. The ratio of the grating separation to the letter size is the same in both patterns.

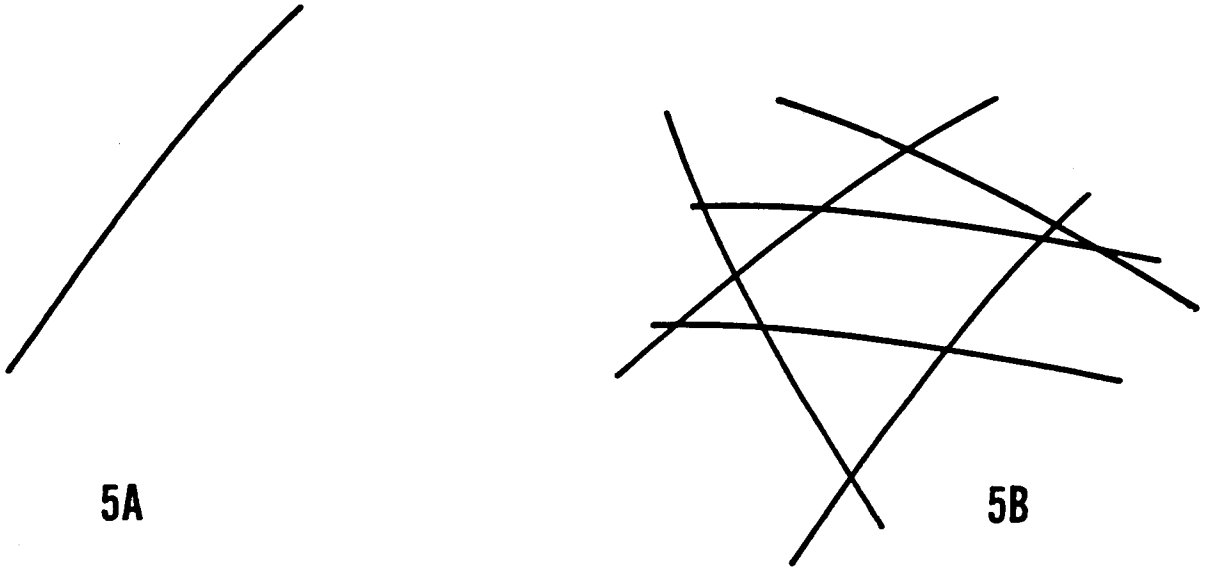

Figure 5

The single line in figure 5a is clearly identifiable with peripheral vision. However the identity of this line is lost when other lines are added to it (figure 5b).

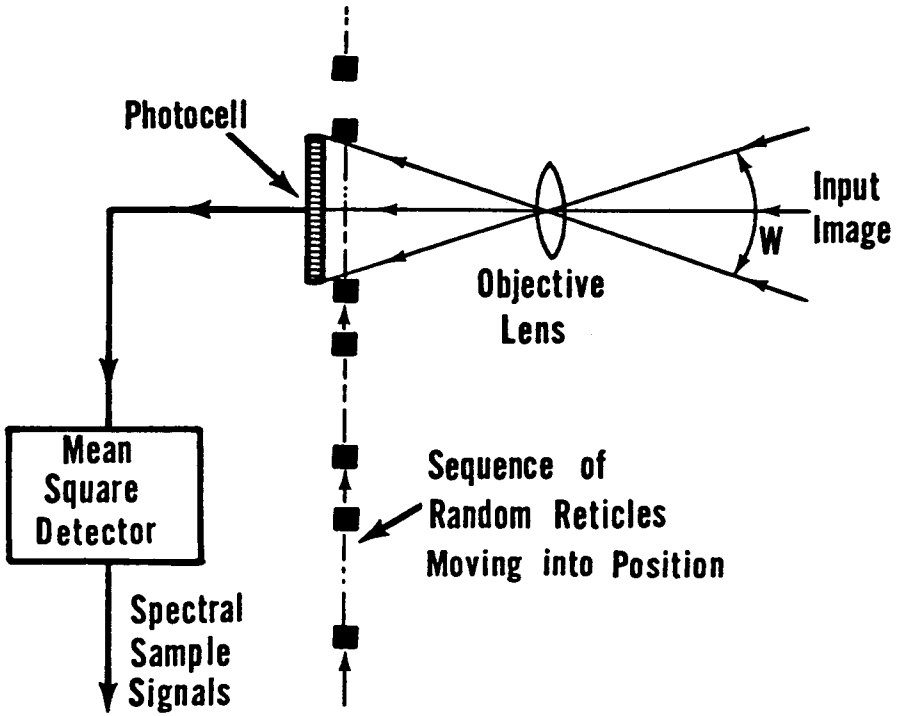

Figure 6

Receptor unit of physical model retina.

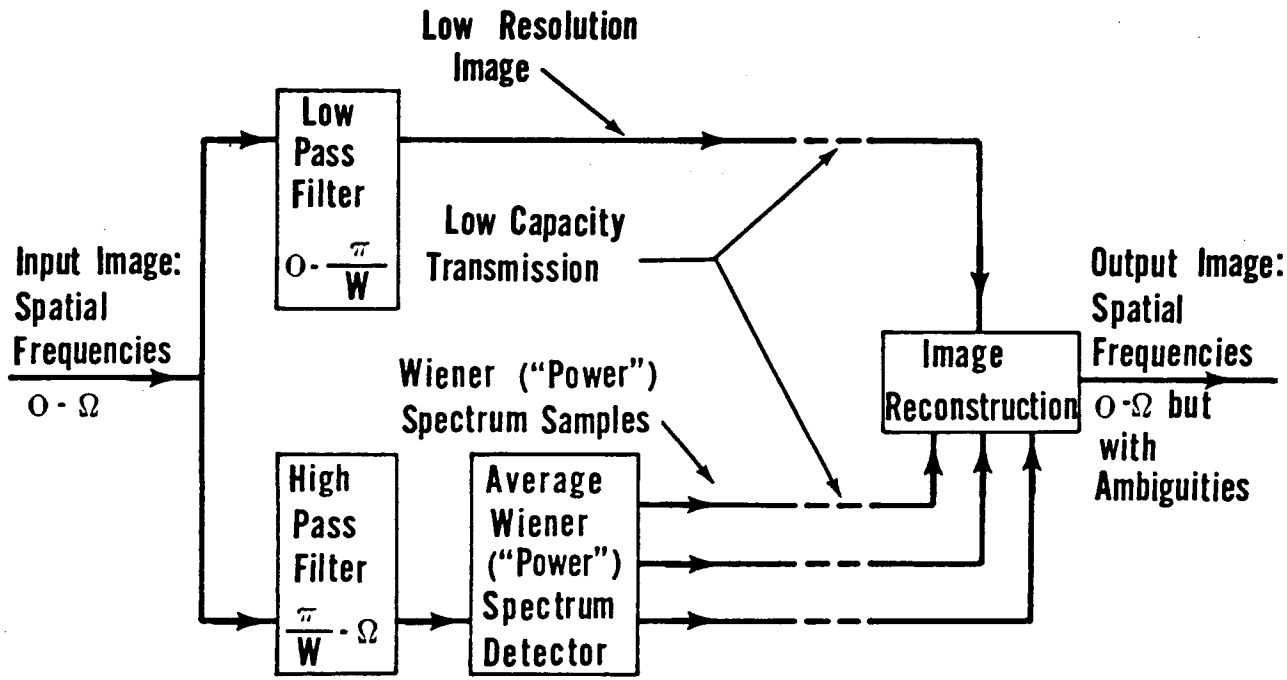

Figure 7

Spatial spectral method of image transmission.

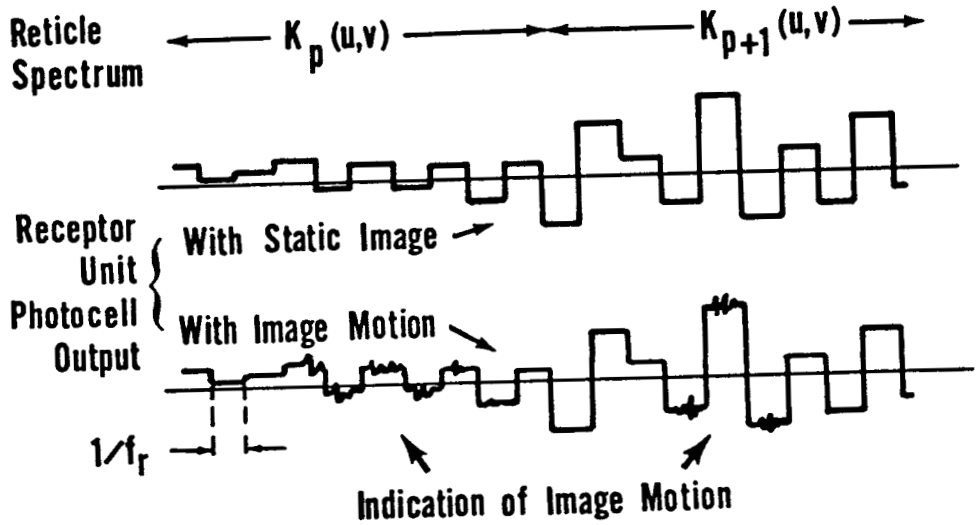

Figure 8

Receptor-unit protocell output as a function of time with and without image motion.